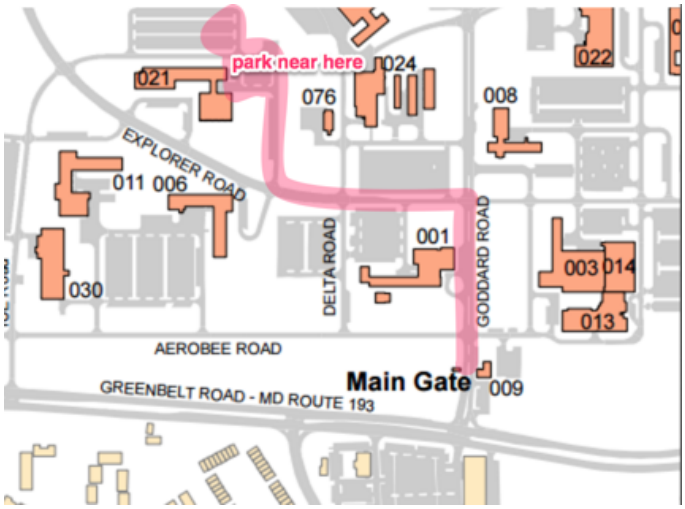
Metadata Curation Summit 1 - Agenda and Meeting Minutes

Where is this meeting? We will be in the library, which is above the cafeteria. Grab your coffee on the way upstairs!



Agenda

February 15, 2017

Time	Topic	Presenter
9:00 - 9:45 AM	Introductions/Icebreaker	Dana
9:45 AM - 10:00 AM	Goals	Kaylin and Katie
10:00 AM - 10:15 AM	Overview of ARC Process	Kaylin
10:15 AM - 10:45 AM	Preview of Dashboard	E84
10:45 AM - 11:05 AM	Deep Dive of ARC Reports and Rule	Kaylin and Jeanne
11:05 AM - 11:15 AM	ARC Schedule	Jeanne
11:15 AM - 11:25 AM	BREAK	
11:25 AM - 11:45 AM	Keywords: Curation vs. Creation Perspectives	Jeanne
11:45 AM - 12:20 PM	Keywords: Change Process	Steve
12:20 PM - 1:15 PM	LUNCH	

1:15 PM - 1:30 PM	CMR Tools	Dana
1:30 PM - 1:45 PM	Temporal Keyword	Kaylin
1:45 PM - 2:00 PM	Spatial Keyword	Kaylin
2:00 PM - 2:15 PM	Spatial Extent	John
2:15 PM - 2:30 PM	Version ID	John
2:30 PM - 2:45 PM	Platform/Sensor/Instrument	John
2:45 PM - 3:00 PM	Citation	Kathy
3:00 PM - 3:15 PM	Processing Level	Kathy
3:15 PM - 3:25 PM	BREAK	
3:25 PM - 4:30 PM	Additional Metadata Topics (DOI, License, Collection Progress, Related URL, Online Resource, Org / Data Contacts, Long Name, Org Names)	Varied
4:30 PM - 4:45 PM	Future Communications	Dana
4:45 PM - 5:00 PM	Action Item Review	Dana

Link to presentations: https://drive.google.com/drive/folders/0BwS_T6Klk_18NGdRbEtlVVhjS2s

Attendees:

- Carey Noll, CDDIS
- Chris Lynnes, EOSDIS
- Katie Baynes, EOSDIS
- John Kozimor, HDF
- Chris Lindsley, ASDC
- Tammy Walker, ORNL
- Jon Pals, Raytheon
- Steve Wharton, GCMD
- Melissa Genazzio, GCMD
- Shannon Leslie, NSIDC
- Simon Cantrell, Raytheon
- Kathy Carr, Raytheon
- Scott Caltagirone, E84
- Andrew Baker, E84
- Merly Hansen, SEDAC
- Dave Meyer, GSFC
- Patrick Williams, ASDC
- Amanda, ASDC
- Erich Reiter, Raytheon
- Leigh Sinclair, GHRC
- Tiffany Matthews, ASDC
- Mary Nair, GHRC
- Jeanne LeRoux, UAH
- Dana Shum, Raytheon
- Tyler Stevens, GCMD
- Kaylin Bugbee, UAH

Goals:

- Katie brought up the FAIR concept from Force11 (https://www.force11.org/group/fairgroup/fairprinciples)
    - Findable
    - Accessible
    - Interoperable
    - Reusable

ARC Process (Kaylin):

- Question: What all should be looked at when evaluating a collection?
- Answer: Metadata itself, html rendering, EDSC and facets

- Question: Will DAACs be expected to fix things w/o questioning it?
- Answer: No! This is a dialogue. Findings marked as "Red" must have something done about them, but that something may be a dialogue to discuss why this isn't a "red"

- Question: Is the process manual or automated
- Answer: Both! Some things are automated, other things are manual.

- Question: How should we handle multi-variable files? What all keywords should be used?
- Answer: ARC team will provide recommendations on which keywords to include which would include not just the variables in the file, but also intended use keywords. Team will also factor in if all variables are searchable.

- Katie Baynes: ARC will distribute all the reports on BEDI records. The DAACs will have a month to respond with a plan and a timeline. The DAAC's timeline should take less than 2 years to achieve

Preview of Dashboard (Scott / Andrew):

- The dashboard is using pyCMR to make calls to the CMR

- Question: Will existing reports be imported into the dashboard?
- Answer: Yes, but in the future.

- [x] Scott Caltagirone - Recommend adding versionID to search screen to help identify collections (this is more appropriate than revision ID)
- [ ] Scott Caltagirone - Recommend using tags to store review status so that all CMR users can access review history
- [ ] Scott Caltagirone - Recommend adding a link from the dashboard into the record in MMT (and perhaps vice-versus) so that users could see their results and quickly fix them
- [ ] Scott Caltagirone - Right now, DAACs may have multiple updates (each it's own revision), but not each of them is ready for a re-review. For now, that requires manual communication between the ARC team and the DAACs. It'd be great if DAACs could set some sort of "ready for re-review" flag/tag.
- [ ] Scott Caltagirone - It'd be great if the DAACs could add field level notes to explain why they did something the way they did (much like the ARC team can add field level notes)

ARC Reports and Schedule (Kaylin / Jeanne')

- Question: Do you check for missing keywords? (Ones that should be there, but aren't)
- Answer: Yes!

- Question: ARC is working on quantifying metadata quality, yet the HDF group was also doing this. As a community, do we want multiple scores (like your credit scores), or convergence on one accepted score? Could the new scoring methodology be used for search relevancy also?
- Answer: No one is sure yet. May be too early to decide. Whatever is decided, we should then have documentation explaining how to get your metadata to score well.

- [x] Kaylin Bugbee - Need to determine how to handle DIF9 records. Consider migrating providers to other formats instead of developing DIF9 curation standards. The goal is to get providers to move off of DIF9.

Keywords: Creation versus Curation (Jeanne')

- [x] Dana Shum - Have Manil and Simon meet to discuss dark data parameter mapper.
- [x] Dana Shum - Future work item - add the ability to MMT to "suggest" keywords based on Manil's dark data tool. Answer: Added EPR-361 to capture this idea for future work.
- [ ] Kaylin Bugbee - Send out the cf <-> gcmd keyword mapping document (or upload to the drive with the other materials)
- [x] Kaylin Bugbee - Consider adding Issues By Field to the dashboard to help determine problem areas.

Keywords: Change Process: (Steve)

- Discussion occurred where it was mentioned that currently GCMD is completing two major keywords updates per year. Steve mentioned the goal of converging towards a major keyword review process and release to an annual process. More widely, as metadata curators, we are striving to coordinate the new UMM release, GCMD keyword release and any ARC recommendations into a yearly event

- Andy Mitchell had requested that all reports on EOSDIS affected records re: GCMD keyword updates be passed through ARC to ensure we account for these changes in our review process.
- Steve Wharton: Recommends that we consider scoring issues by field with the number of issues being displayed on the y-axis and the field names being displayed on the x-axis.
- Question: Can you add https entries on docBuilder?
- Answer: Yes!

Temporal Keywords:

- Noted that this field is probably not used in search
- When there is an obvious mapping to the provided GCMD keywords, ARC should make the recommendation to the DAAC to map to the GCMD keywords.
- When there is not an easy or obvious mapping and/or if the provided information is not adding clarity or value to the record, ARC should recommend removing the temporal keyword since it is not a required element.
- DAACs should standardize the usage of terms within the organization. This will at least guarantee some consistency within each DAAC and may be the best that we can hope for since it is a free text, uncontrolled field.
- There did not seem to be a desire from the group to take a middle path and work towards developing either:
    - A best practices document or guidelines document for providing temporal information for field campaign and airborne data.
    - Removing the temporal keyword element and adding 2 new temporal elements that are more rigorously defined (File Temporal Range and Sampling Frequency)

- Temporal keywords are very tricky, as it could mean the span of the collection, the span of each file and/or the sampling frequency. The metadata makes this hard to capture today
- Recommend keeping the collection level as "varies" for the complex cases and allowing the granules/files to further define it.

- [x] stephen wharton - Recommend scrubbing the temporal keywords in GCMD to better align with the communities needs, then we could standardize and have this as controlled vocabulary
- [x] Dana Shum - Validate how the CMR searches collections where there are multiple temporal extents which are discontinuous (as in, from different flights). Right now, it looks like we search the full range, versus each discrete range and we need to correct that. Answer: Today it uses the earliest and latest elements of any extent. I filed CMR-4103 to fix this. We don't believe there are any technical challenges with doing so, this was just different than the initial requirements.

Location/Spatial Keywords:

- Dana: NLP can extract some information about spatial names. Perhaps some spatial keywords like country, state, etc… are now unnecessary since NLP can identify this information especially if CMR is using NLP results before spatial keywords results

- Question: Where does GCMD get the list of keywords?
- Answer: GCMD currently maps the keywords to ISO standards to ensure a broad, standards-based listing

- Question: Problems arise when teams want to use keywords for things like various basins or localized names. How should we as a community handle these?
- Answer: Add keywords as needed.

- Shannon Leslie: Having spatial keywords that describe things like basins, watersheds, etc… may be more beneficial for describing metadata since these locations may not be easily identified from NLP and other CMR tools. This requires work in partnering with the GCMD to expand the spatial keyword list to include this new information.
- Discussion regarding the possibility of providing a detailed controlled field that you can append to a GCMD hierarchy list. Tyler

mentioned that if an added detailed controlled field had a lot of use, it could be added to the controlled list. There was some concern expressed about giving the providers the ability to add whatever they wanted without picking from a controlled vocabulary. Tiffany made an excellent point that left to their own devices users will create many iterations of a 'keyword' that essentially states the same thing. This is why a controlled vocabulary is important – it ensures consistency. There needs to be a consensus built between CMR, ARC and GCMD on controlled vocabularies and adding 'one off' keywords. Is there an action here to schedule a discussion?
- If the DAACs are going to request spatial keyword updates that are unique, the GCMD will need to be more responsive than updating keywords annually. Perhaps this is ok since spatial keywords may not need to be subjected to subject matter review.

- [x] Dana Shum - Ensure location keywords are indexed. Answer: They are.
- [x] Dana Shum - Work with ESDIS to determine an approach for "locking down" the lowest level of the keywords. If the lowest level is uncontrolled, then we have no way to ensure consistency in the data.: Answer: In talking this through with ARC, the ARC team will ensure that EOSDIS DAACs are using controlled vocabulary and adding any missing keywords to the controlled list. This should ensure that all EOSDIS DAACs are only using controlled vocabulary. At some point before the CMR begins "enforcing" keyword compliance, we will then need to ensure that IDN records use controlled vocabulary also, but that is in the future.

Instrument/Sensor:

- People like the new approach to simplify this part of the metadata using the "composed-of" relationship
- People expressed a desire to nest sensors beyond one level
- In order to make the Platform/Instrument hierarchy effective, reconciliation will need to happen. There are 149 unique platform names in CMR that do not map to GCMD and 255 instrument names that do not map to GCMD. This will be an important part of the ARC activity.
- When providers run into places where their instrument doesn't exist as a keyword, they should reach out to GCMD to add it

- [ ] Dana Shum / Kaylin Bugbee- Recommend adding validation on Platform -> Instrument relationships. *Status: Raising this question at Curation Summit 2.*
- [x] Dana Shum - Determine if we could better utilize the category of an instrument (on facets). Answer: Added CMR-4104 to make this change.

VersionID:

- "Processing Level" may not be a good term for model output
- Missing "What's new that drove the new version"
- Implementing this would need to be a long term, top down drive
- Many times, Version ID is driven by what the PI wants, so metadata authors would need ESDIS "hammer" to help make it official

- [x] John Farley - consider changing "local Version ID" to "Producer Version ID". Answer: Updated proposal for Curation Summit 2
- [ ] John Farley - we should bring this topic to the ESIP Data Stewardship cluster after the next Curation Summit. It'd be a good place to get broader consensus.

Spatial Extent:

- Concurrence on needed multiple bounding boxes
- Group agreed that the two recommendations should be done: (Incorporate ephemeris info and eliminate ambiguities in spatial representation)
- Group agreed that the study could be delayed (or tabled if needed) as there was not a strong need for it right now

Citation:

- Many DAACs are recommending sites like crosssite.org where you can populate a DOI and have it generate the citation for you
- SEDAC also has a great citation manager tool

- Question: Could remove citation all together and just have people use DOI to construct the citation?
- Answer: No, we need to remember that the IDN providers do not all have DOIs, so this would not work

- [ ] Kathleen Carr - Group consensus is that option 1 is the best option. De-duplicate the UMM-C to simplify it and then add the individual blocks to ECHO10 format. Need to create tickets for this work. *Status: Confirming at Curation Summit 2 before acting*.
- [ ] Kathleen Carr - Need to assess what to display on EDSC (perhaps link to something like crosssite.org and pass in requested info. *Status : Confirming at Curation Summit 2 before acting*

Processing Level:

- Recommendation is to humanize values to the NASA approved processing levels and also include a short description ("Raw (0)", etc.) so that the facets are concise and meaningful
- We received some push back from a few representatives about changing processing level information to conform with the NASA processing levels.
- Tammy Walker, ORNl said they went through and made everything NASA compliant. Said it wasn't a difficult task.
- 'Not Provided' was suggested as an entry. Not entirely comfortable with that label since processing level information is required.
- Descriptions for each level are needed and should be standardized.

- [x] Dana Shum - assign someone to write-up the processing level humanizers: Answer: John Farley is working this now.
- [ ] Kathleen Carr - Need to ensure "not provided" doesn't show up in the facets
- [x] Kathleen Carr - Validate that once humanized, we can search on the original value and the humanized values. Answer: They are.
- [ ] Kaylin Bugbee - ARC team should recommend providers use the NASA standards.

DOI:

- No concerns with what was discussed in the slides

Licenses:

- [ ] Kathleen Baynes - Need ESDIS to determine what license should be used. Most likely, the correct license for NASA Earth Science data is 'U.S. Public Domain'. Need to also check on international implications.

- This field will most likely be required for EOSDIS DAACs, but not for all CMR.

Collection Progress:

- Currently seem to be co-mingling progress with maturity state. Recommend splitting into two:
  - Progress: Complete, Active, Planned
  - Maturity: Beta, Provisional, Validated, Deprecated

- [ ] Dana Shum - Work with ESDIS and GCMD to determine if small lists like this should be controlled in KMS or simply in a schema *Status : Posed this as a question for MetadataCurationSummit 2*

General:

- [ ] Kaylin Bugbee - Consider doing an ESO review of the curation rules to ensure wide acceptance of the rules
- [x] Kaylin Bugbee / Dana Shum - Consider a metadata authoring forum or FAQ. If it's a forum, we would need a way to mark answers as "final" so that people know when a recommendation is the official answer. — Answer: Team will use #curation on slack for questions and debate and then start a FAQ on this wiki space when we have consensus.
- [ ] Christopher Lynnes - Build guidebook for data producers (when to use NetCDF4, guidelines to follow, etc.)
- [x] Kaylin Bugbee - DAACs request that when ARC sends out a report, they give the DAACs time to digest it and then setup a follow-up meeting a week or two after to answer immediate questions.
- [x] Tyler Stevens - Respond to the recommendation to review and add the additional URL Content types.
- [x] Tyler Stevens - Respond to the recommendation to add an element to GCMD's provider keywords for "Abbreviation" so that providers can specify a better short name for their provider.

- Question: Does bulk update of granules exist?
- Answer: No, not yet. But if people want it, we would really like to build it.